Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement

V. Kuentz Simonet, Sandrine Lyser, Jacqueline Candau, Philippe Deuffic, Marie Chavent, Jérôme Saracco

Journal de la Société Française de Statistique
Vol. 154 No. 2 (2013)

Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement

Title: A variable clustering approach for the typology of units: a survey on farming and environment

Vanessa Kuentz-Simonet[1], Sandrine Lyser[1], Jacqueline Candau[1], Philippe Deuffic[1]
Marie Chavent[2] et Jérôme Saracco[2]

Résumé : Nous considérons le cas d'une enquête agriculture/environnement dont les données sont relatives aux transformations actuelles du métier d'agriculteur. Nous optons pour une démarche originale en remplaçant la première étape classique d'analyse factorielle par un algorithme de classification de variables. L'objectif de la classification de variables est de construire des classes de variables fortement liées entre elles et de supprimer ainsi l'information redondante. L'approche ClustOfVar utilisée fournit simultanément des groupes de variables ainsi que les variables synthétiques associées aux classes de variables. Dans cet algorithme, le critère d'homogénéité repose sur la notion de corrélation pour les variables quantitatives et de rapport de corrélation pour les variables qualitatives. L'étape de classification de variables nous permet d'obtenir des variables synthétiques que nous proposons de lire comme une sorte de gradient. Sur nos données, les valeurs correspondent à des regroupements de modalités distincts et pertinents pour l'interprétation. Cette démarche nous permet de lire et d'étiqueter chaque variable synthétique. Nous mettons ainsi en évidence des tendances qui vont départager l'opinion des agriculteurs quant à leur prise en compte de l'environnement. Puis nous précisons ces résultats en réalisant une classification sur les scores des individus mesurés sur les variables synthétiques. Sur le plan sociologique, l'apport des variables synthétiques pour interpréter les profils-types obtenus est incontestable.

Abstract: A survey on farming and environment dealing with the current transformations of the farmer job is considered. We propose to replace the usual data mining strategy which consists of applying Multiple Correspondence Analysis by a variable clustering approach. Clustering of variables aims at lumping together variables which are strongly related to each other and thus bring the same information. The ClustOfVar approach used in this paper provides at the same time groups of variables and their associated synthetic variables. In this algorithm, the homogeneity criterion of a cluster is defined by the squared Pearson correlation for the quantitative variables and by the correlation ratio for the qualitative variables. The step of variable clustering enables to get synthetic variables that can be read as a gradient. In our case study, values correspond to some relevant groupings of categories. This enables to interpret and name easily the synthetic variables. Trends in the opinion of farmers are thus highlighted with the variable clustering approach. Then we clarify these first results by applying a clustering method on the scores of the individuals measured by the synthetic variables. At the sociological level, the supply provided by the synthetic variables to interpret the clusters of farmers is obvious.

Mots-clés : classification de variables, variables synthétiques, typologie d'agriculteurs, environnement
Keywords: variable clustering, synthetic variables, typology of farmers, environment
Classification AMS 2000 : 62-07, 62H99, 62P12

[1] Irstea, UR ADBX
E-mail : vanessa.kuentz-simonet@irstea.fr and E-mail : sandrine.lyser@irstea.fr and E-mail : jacqueline.candau@irstea.fr and E-mail : philippe.deuffic@irstea.fr
[2] INRIA Bordeaux Sud-Ouest, Équipe CQFD
E-mail : marie.chavent@math.u-bordeaux1.fr and E-mail : jerome.saracco@math.u-bordeaux1.fr

 ISSN: 2102-6238

1. Introduction

L'impératif environnemental joue aujourd'hui un rôle dans la recomposition des identités professionnelles agricoles. On peut cependant se demander si pour les agriculteurs, cet impératif est prédominant ou si d'autres facteurs contextuels ne viennent pas également modifier leur conception du métier. Nous examinons cette question grâce à une enquête quantitative qui a été confiée à Irstea en 2005 par le Centre national pour l'aménagement et les structures des exploitations agricoles, organisme public chargé de la gestion financière des aides publiques attribuées aux exploitants agricoles. La prise en compte de l'environnement n'est pas une variable binaire, contrairement à la contractualisation des mesures agri-environnementales ou l'adhésion à un réseau d'agriculture alternative par exemple, séparant les agriculteurs qui intègrent les préoccupations environnementales et les autres. Dès lors, l'objectif de l'étude s'est centré sur la façon dont les agriculteurs conçoivent la protection de l'environnement en relation avec leur activité : quelles significations attribuent-ils à l'environnement ? Quelles valeurs et dimensions de leur métier et de leur rapport à la nature sont remises en cause par l'inscription de la protection de l'environnement dans la politique agricole ?

Les méthodes d'analyses multidimensionnelles sont alors privilégiées pour répondre à cette problématique. En statistique exploratoire multidimensionnelle, la classification des observations est couramment utilisée pour établir des profils-types. Une stratégie classique consiste à réaliser une analyse factorielle des données puis à appliquer une méthode de classification sur les scores des individus (mesurés sur les composantes principales obtenues). D'autres alternatives ont été proposées pour réaliser simultanément la réduction du nombre de variables et la classification des observations. De Soete and Carroll (1994) introduisent une méthode appelée "k-means clustering procedure in a reduced space" qui est basée sur le critère défini dans l'algorithme des k-means. L'idée est d'optimiser ce critère sous la contrainte que les centres des classes appartiennent à un sous-espace engendré par les colonnes de la matrice des données. Vichi and Kiers (2001) proposent une approche nommée "factorial k-means" afin de déterminer un sous-espace de représentation des données tel que les points projetés aient la plus petite distance aux centres des classes. Cette approche, combinant algorithme des k-means et ACP, permet de sélectionner les composantes les plus pertinentes pour la classification en minimisant un seul et même critère. Les auteurs précisent que leur approche comporte un inconvénient lorsque les données présentent des dimensions avec des variances faibles : elle se concentre en priorité sur ces dimensions dans la mesure où elles contribuent peu à la fonction de perte. Pour surmonter ce problème, les auteurs préconisent de supprimer au départ les dimensions triviales présentes dans les données. La description de cette méthode ainsi qu'une comparaison avec la méthode "factorial k-means" est réalisée dans Timmerman et al. (2010). D'autres approches combinant une méthode de classification et la recherche d'un sous-espace de représentation ont été proposées. On peut citer les travaux relatifs au "multidimensional scaling" ou "unfolding analysis" Heiser (1993) ; De Soete and Heiser (1993) ; DeSarbo et al. (1991)) ou plus récemment, l'approche par modèle de mélange de Govaert and Nadif (2009) pour la classification croisée et les travaux de Vichi and Saporta (2009) intitulés "clustering and disjoint principal component analysis". Le lecteur peut aussi se référer à l'article de Charrad and Ben Ahmed (2011) qui présente une revue des méthodes de bi-partitionnement. Cependant nombre de ces approches sont dédiées à l'analyse de données quantitatives. À notre connaissance, le cas des données qualitatives ou mixtes (mélange de données quantitatives et

 ISSN: 2102-6238

qualitatives) a reçu moins d'attention.

Dans cet article, nous optons pour une approche différente, qui sera par ailleurs valable pour des données quantitatives, qualitatives ou mixtes. Plus précisément, nous proposons de remplacer la première étape d'analyse factorielle par une approche de classification de variables que nous avons récemment développée (Chavent et al. (2011), Chavent et al. (2012a)). L'objectif est de supprimer dans un premier temps l'information redondante et de réduire ainsi la dimension du tableau. Plus précisément, en réorganisant les variables en classes homogènes, l'approche de classification construit simultanément des variables synthétiques. Notons que quel que soit le type des données initiales, les variables synthétiques construites sont toujours quantitatives et peuvent être lues comme une sorte de gradient. Sur nos données, la lecture de ces variables synthétiques fournit des premiers éléments d'interprétation intéressants. L'étape suivante consiste à classifier les individus à partir de leurs scores observés sur les variables synthétiques.

L'approche par classification de variables est présentée dans la section 2. La démarche méthodologique est illustrée dans la section 3 à l'aide d'une étude de cas relative à la prise en compte de l'environnement par les agriculteurs français. Notons que les données de cette application sont de nature qualitative. Cependant nous attirons l'attention du lecteur sur le fait que nous avons choisi de décrire la démarche dans sa généralité. Ainsi l'approche de classification de variables de la section 2 est présentée pour un ensemble de variables quantitatives et/ou qualitatives. Pour finir, la section 4 discute la pertinence de la démarche et fournit des éléments de conclusion.

2. Une approche par classification de variables

2.1. Introduction sur la classification de variables

L'objectif de la classification de variables est de regrouper entre elles des variables liées, c'est-à-dire porteuses de la même information, afin de construire des classes de variables homogènes (le sens de ce terme sera précisé plus loin). Dans de nombreuses applications, on s'intéresse à la classification des variables et non à celle des individus. C'est le cas par exemple en analyse sensorielle (mise en place de groupes de descripteurs), en biochimie (classification de gènes), en marketing (segmentation d'un panel de consommateurs), en économie (détection de stratégies financières), etc. On peut citer par exemple les travaux de Plasse et al. (2007) qui utilisent la classification de variables dans la recherche de règles d'association pour une application issue de l'industrie automobile. Un autre objectif poursuivi par la classification de variables est la suppression des redondances entre les variables et ainsi la réduction de la dimension du tableau de données. Dans ce cas, après avoir construit des groupes de variables liées, il est nécessaire de sélectionner dans chaque classe une variable ou de résumer chaque classe de variables par une variable synthétique. La classification de variables apparaît alors comme une alternative aux méthodes d'analyse factorielle classiques que sont l'Analyse en Composantes Principales (ACP) ou l'Analyse des Correspondances Multiples (ACM).

Une approche simple et couramment utilisée pour classifier un ensemble de variables consiste à calculer une matrice de dissimilarités entre les variables puis à appliquer une méthode usuelle de classification dédiée à l'origine aux individus. Pour les variables quantitatives, de nombreuses mesures de dissimilarités peuvent être utilisées. Elles font intervenir par exemple le coefficient de corrélation, la mesure d'association de Soffritti (1999), la distance basée sur l'opérateur d'Es-

 ISSN: 2102-6238

coufier, etc. Concernant les variables qualitatives, le nombre de critères d'association disponibles est tout aussi important : χ^2, Rand, Belson, Jordan, etc. (voir par exemple Abdallah and Saporta (1998) ou Derquenne (2001)). On peut également citer les travaux de Qannari et al. (1998) qui proposent une distance euclidienne entre variables quantitatives permettant de tenir compte aussi bien des variances des variables que de leurs corrélations. Les auteurs étendent cette distance au cas de variables qualitatives et pour un mélange de données quantitatives et qualitatives.

Parallèlement à ce type d'approches, des méthodes ont été spécifiquement développées pour la classification de variables. Pour des données quantitatives, on peut citer entre autres l'approche de Hastie et al. (2000) en biologie génomique. Mais la plus célèbre est sans doute la fonction VARCLUS du logiciel SAS. Cette procédure complexe avec peu de justifications théoriques quant aux options offertes fournit une hiérarchie ou une partition des variables quantitatives. Une autre approche consiste à utiliser un algorithme de classification qui fournit simultanément des classes de variables et leurs variables synthétiques. Deux algorithmes de partitionnement de ce type existent déjà pour la classification de variables quantitatives et sont basés sur l'ACP : la méthode Clustering of variables around Latent Variables (CLV) proposée dans Vigneau and Qannari (2003) et Vigneau et al. (2006) et la méthode Diametrical Clustering développée par Dhillon et al. (2003). À notre connaissance, la classification de variables qualitatives a cependant reçu moins d'attention. On peut citer entre autres l'Analyse de la Vraisemblance du Lien de Lerman (Lerman (1990), Lerman (1993)) qui permet de réaliser une classification hiérarchique de variables quantitatives ou qualitatives.

Nous avons récemment proposé une méthode spécifiquement dédiée à la classification de variables (ClustOfVar) quel que soit leur type, quantitatif, qualitatif ou un mélange des deux (Chavent et al. (2012a)). Cette approche généralise la méthode CLV au cas de données mixtes. Dans CLV, le critère d'homogénéité d'une classe de variables quantitatives est défini comme la somme des corrélations au carré des variables de la classe à la variable synthétique. Cette dernière est obtenue grâce à une ACP des variables de la classe. Dans ClustOfVar, nous étendons le critère d'homogénéité au cas de données mixtes et nous utilisons la méthode PCAMIX de Kiers (1991). Pour cela, nous avons proposé une réécriture de PCAMIX sous forme d'une décomposition en valeurs singulières (Chavent et al. (2012b)). Plus précisément, PCAMIX est une méthode d'analyse en composantes principales pour un mélange de variables quantitatives et qualitatives, qui inclut l'ACP et l'ACM comme cas particuliers. Dans ClustOfVar, deux algorithmes de classification de variables sont proposés : un algorithme ascendant hiérarchique et un algorithme de partitionnement de type k-means. Ces algorithmes visent à maximiser un critère d'homogénéité, basé sur le carré de la corrélation de Pearson pour des variables quantitatives et sur le rapport de corrélation pour des variables qualitatives.

Dans cet article, nous utilisons l'approche ClustOfVar de classification de variables comme alternative à l'analyse factorielle en tant qu'étape préliminaire à la typologie d'observations. N'ayant pas d'idée a priori sur le nombre de classes de variables, seul l'algorithme de classification ascendante hiérarchique sera décrit pour présenter la démarche méthodologique. Notons que l'implémentation des algorithmes de classification de variables est disponible dans le package R nommé ClustOfVar. Pour plus de détails sur l'algorithme de partitionnement ou sur les fonctionnalités du package, le lecteur peut se référer à Chavent et al. (2012a).

 ISSN: 2102-6238

2.2. Notations

Soient $\{\mathbf{x}_1,\ldots,\mathbf{x}_{J^1}\}$ un ensemble de J^1 variables quantitatives et $\{\mathbf{z}_1,\ldots,\mathbf{z}_{J^2}\}$ un ensemble de J^2 variables qualitatives. Notons $\mathbf{X}$ et $\mathbf{Z}$ les matrices des données quantitatives et qualitatives correspondantes, de dimension respective $I\times J^1$ et $I\times J^2$, où I est le nombre d'observations. Dans un souci de simplicité, nous notons $\mathbf{x}_j\in\mathbb{R}^I$ la j-ème colonne de $\mathbf{X}$ et $\mathbf{z}_j\in(\mathscr{M}_j)^I$ la j-ème colonne de $\mathbf{Z}$ avec $\mathscr{M}_j$ l'ensemble des modalités de $\mathbf{z}_j$. Enfin $\mathscr{P}_K=(C_1,\ldots,C_K)$ désigne une partition de l'ensemble des $J=J^1+J^2$ variables en K classes.

2.3. Homogénéité $\mathscr{H}$ d'une partition $\mathscr{P}_K$ de variables

L'homogénéité d'une partition $\mathscr{P}_K$ de l'ensemble des variables est définie par la somme des homogénéités des classes qui la composent :

$$\mathscr{H}(\mathscr{P}_K)=\sum_{k=1}^{K}H(C_k), \tag{1}$$

où $H(C_k)$ mesure l'homogénéité de la classe C_k de $\mathscr{P}_K$. Il s'agit d'une mesure d'adéquation entre les variables de la classe et la variable synthétique quantitative de la classe, notée $\mathbf{y}_k\in\mathbb{R}^I$:

$$H(C_k)=\sum_{\mathbf{x}_j\in C_k}r^2_{\mathbf{y}_k,\mathbf{x}_j}+\sum_{\mathbf{z}_j\in C_k}\eta^2_{\mathbf{y}_k|\mathbf{z}_j}, \tag{2}$$

où r^2 désigne la corrélation de Pearson au carré et η^2 désigne le rapport de corrélation. Plus précisément le rapport de corrélation $\eta^2_{\mathbf{y}_k|\mathbf{z}_j}\in[0,1]$ mesure le pourcentage de variance de $\mathbf{y}_k$ expliquée par les modalités de $\mathbf{z}_j$:

$$\eta^2_{\mathbf{y}_k|\mathbf{z}_j}=\frac{\sum_{s\in\mathscr{M}_j}I_s(\bar{y}_k^s-\bar{y}_k)^2}{\sum_{i=1}^{I}(y_{ik}-\bar{y}_k)^2},$$

où I_s désigne l'effectif de la modalité s, $\bar{y}_k^s$ est la valeur moyenne de $\mathbf{y}_k$ calculée sur les observations possédant la modalité s, y_{ik} est la valeur de $\mathbf{y}_k$ pour l'observation i, et $\bar{y}_k$ est la moyenne de $\mathbf{y}_k$.

Notons que le premier terme de $H(C_k)$ (utilisant la corrélation au carré r^2) mesure le lien entre les variables quantitatives de C_k et $\mathbf{y}_k$, indépendamment du signe de la relation. Le second terme (avec le rapport de corrélation η^2) mesure le lien entre les variables qualitatives de C_k et la variable synthétique $\mathbf{y}_k$. Ainsi l'homogénéité d'une classe est maximale lorsque toutes les variables quantitatives sont parfaitement linéairement corrélées (positivement ou négativement) avec $\mathbf{y}_k$ et lorsque tous les rapports de corrélation des variables qualitatives à $\mathbf{y}_k$ sont égaux à 1.

2.4. Définition de la variable synthétique d'une classe de variables C_k

La variable synthétique quantitative $\mathbf{y}_k\in\mathbb{R}^I$ de la classe C_k est définie comme la variable "la plus liée" (en un certain sens) aux variables de la classe. Elle maximise l'homogénéité de la classe C_k et est donc solution du problème d'optimisation suivant :

$$\mathbf{y}_k=\arg\max_{\mathbf{u}\in\mathbb{R}^I}\left\{\sum_{\mathbf{x}_j\in C_k}r^2_{\mathbf{u},\mathbf{x}_j}+\sum_{\mathbf{z}_j\in C_k}\eta^2_{\mathbf{u}|\mathbf{z}_j}\right\}.$$

On peut montrer que :

 ISSN: 2102-6238

– $\mathbf{y}_k$ est la première composante principale issue de PCAMIX appliqué à $\mathbf{X}_k$ et $\mathbf{Z}_k$, les matrices formées par les colonnes de $\mathbf{X}$ et $\mathbf{Z}$ correspondant aux variables présentes dans la classe C_k ;
– la variance empirique de $\mathbf{y}_k$ vaut alors : $\mathrm{Var}(\mathbf{y}_k) = \sum_{\mathbf{x}_j \in C_k} r^2_{\mathbf{y}_k,\mathbf{x}_j} + \sum_{\mathbf{z}_j \in C_k} \eta^2_{\mathbf{y}_k|\mathbf{z}_j} = \lambda^1_k$, la première valeur propre issue de PCAMIX appliquée à la classe C_k.

Il en découle que l'homogénéité d'une classe est simplement définie par $H(C_k) = \lambda^1_k$. De ce fait, l'homogénéité de la partition $\mathscr{P}_K$ est égale à $\mathscr{H}(\mathscr{P}_K) = \lambda^1_1 + \cdots + \lambda^1_K$.

2.5. *Calcul de la variable synthétique d'une classe de variables C_k en utilisant PCAMIX*

Nous utilisons une présentation de PCAMIX sous forme d'une décomposition en valeurs singulières (voir Chavent et al. (2012b) pour plus de détails sur cette méthode). Notons que cette méthode est implémentée dans un package R nommé PCAmixdata. Il sera prochainement accessible sur le CRAN et est actuellement disponible auprès des auteurs. Le calcul de $\mathbf{y}_k$ se fait selon les étapes suivantes :

1. Recodage de $\mathbf{X}_k$ et $\mathbf{Z}_k$:
   (a) $\tilde{\mathbf{X}}_k$ est la version standardisée de la matrice des données quantitatives $\mathbf{X}_k$.
   (b) On note $\mathbf{I}_I - \mathbf{1}\mathbf{1}'/I$ l'opérateur de centrage, avec $\mathbf{I}_I$ la matrice identité d'ordre I et $\mathbf{1}$ le vecteur colonne de $\mathbb{R}^I$ composé de 1. La matrice $\tilde{\mathbf{Z}}_k = (\mathbf{I}_I - \mathbf{1}\mathbf{1}'/I)\mathbf{G}\mathbf{D}^{-1/2}$ désigne alors la version standardisée du tableau disjonctif complet $\mathbf{G}$ de la matrice $\mathbf{Z}_k$ des données qualitatives. $\mathbf{D}$ est la matrice diagonale des fréquences relatives des modalités des variables de la classe. On a $\mathbf{D} = \mathrm{diag}(\frac{I_s}{I}), s = 1, \ldots, m_k$ avec m_k le nombre total de modalités des variables qualitatives dans C_k.
2. Concaténation des deux matrices recodées : $\mathbf{M}_k = \frac{1}{\sqrt{I}}(\tilde{\mathbf{X}}_k|\tilde{\mathbf{Z}}_k)$.
3. Décomposition en valeurs singulières de $\mathbf{M}_k$: $\mathbf{M}_k = \mathbf{U}_k\Lambda_k\mathbf{V}'_k$ où $\mathbf{U}'_k\mathbf{U}_k = \mathbf{V}'_k\mathbf{V}_k = \mathbf{I}_r$ où r est le rang de $\mathbf{M}_k$, et Λ_k est la matrice diagonale des valeurs singulières $\lambda^1_k, \ldots, \lambda^r_k$ rangées par ordre décroissant.
4. Calcul de la matrice des scores des composantes principales de dimension $I \times r$: $\sqrt{I}\mathbf{U}_k\Lambda_k$.
5. Extraction de la variable synthétique $\mathbf{y}_k$ (première colonne de cette matrice) :

$$\mathbf{y}_k = \sqrt{I}\lambda^1_k\mathbf{u}^1_k, \tag{3}$$

   où $\mathbf{u}^1_k$ désigne la première colonne de $\mathbf{U}_k$.

Remarque : La matrice $\mathbf{V}_k$ des vecteurs propres contient les coefficients de la combinaison linéaire des variables dans l'expression des composantes principales.

Le lecteur peut se référer à l'Annexe 1 pour une illustration du calcul de la variable synthétique d'une classe avec PCAMIX.

2.6. *Une matrice essentielle dans les résultats de PCAMIX pour l'interprétation des résultats*

Suite au calcul de la variable synthétique de chaque classe de variables C_k avec PCAMIX, nous pouvons également calculer la matrice $\mathbf{A}_k$ suivante :

$$\mathbf{A}_k = \mathbf{V}_k\Lambda_k. \tag{4}$$

 ISSN: 2102-6238

Pour interpréter plus aisément les valeurs de cette matrice, nous écrivons $\mathbf{A}_k$ comme la concaténation de deux matrices $\mathbf{A}_k^1$ et $\mathbf{A}_k^2$ qui font respectivement référence aux variables quantitatives et qualitatives de la classe C_k. Plus précisément $\mathbf{A}_k^1$ est de dimension $J_k^1 \times r$, avec J_k^1 le nombre de variables quantitatives dans C_k et $\mathbf{A}_k^2$ est de dimension $m_k \times r$, avec m_k le nombre total de modalités des variables qualitatives dans C_k. On note donc :

$$\mathbf{A}_k = \begin{pmatrix} \mathbf{A}_k^1 \\ \mathbf{A}_k^2 \end{pmatrix}, \tag{5}$$

où :

- la matrice $\mathbf{A}_k^1$ contient les corrélations des J_k^1 variables quantitatives de la classe C_k avec les r composantes principales issues de PCAMIX.
- la matrice $\mathbf{D}^{-1/2}\mathbf{A}_k^2$ contient les coordonnées des m_k modalités des J_k^2 variables qualitatives de la classe C_k sur les composantes principales.

On peut également calculer la matrice $\mathbf{C}_k$ qui est une sortie classique des méthodes d'analyse factorielle, très utile en présence de données mixtes. La matrice $\mathbf{C}_k = (c_{jl}), j = 1,\ldots,J_k^1+J_k^2; l = 1,\ldots,r$ est obtenue à partir de $\mathbf{A}_k$ de la façon suivante :

$$c_{jl} = \begin{cases} a_{jl}^2 & \text{si la variable } j \text{ est quantitative,} \\ \sum_{s\in I_j} a_{sl}^2 & \text{si la variable } j \text{ est qualitative,} \end{cases}$$

où I_j désigne l'ensemble des indices des lignes de $\mathbf{A}_k$ associées aux modalités de la variable qualitative j. Dans le cas des variables quantitatives, c_{jl} est la corrélation au carré entre la variable et la composante principale. Pour les variables qualitatives, on peut montrer qu'il s'agit du rapport de corrélation de la variable avec la composante principale.

Nous verrons dans la sous-section 2.8 que ces matrices $\mathbf{A}_k^1$, $\mathbf{D}^{-1/2}\mathbf{A}_k^2$ et $\mathbf{C}_k$ jouent un rôle essentiel dans l'interprétation des variables synthétiques.

2.7. L'algorithme de classification ascendante hiérarchique de variables

L'objectif est de trouver une partition de l'ensemble des variables quantitatives et qualitatives telle que les variables à l'intérieur d'une classe soient fortement liées entre elles. Il s'agit de maximiser le critère d'homogénéité $\mathscr{H}$ défini dans (1). Pour cela, un algorithme de classification ascendante hiérarchique est proposé. Il construit un ensemble de J partitions emboîtées de variables de la façon suivante :

1. Étape $l = 0$: initialisation avec la partition des singletons (J classes).
2. Étape $l = 1,\ldots,J-2$: agrégation de deux classes de la partition en $J-l+1$ classes pour obtenir la nouvelle partition en $J-l$ classes. Pour cela, on agrège les deux classes A et B qui ont la plus petite dissimilarité d définie par :

$$d(A,B) = H(A)+H(B)-H(A\cup B) = \lambda_A^1+\lambda_B^1-\lambda_{A\cup B}^1. \tag{6}$$

On peut montrer que $\lambda_{A\cup B}^1 \leq \lambda_A^1+\lambda_B^1$, ce qui implique que l'agrégation de deux classes A et B entraîne une baisse du critère $\mathscr{H}$. Cette dissimilarité mesure donc la perte en homogénéité observée quand les deux classes A and B sont agrégées. La stratégie consiste

 ISSN: 2102-6238

donc à agréger les deux classes qui entraînent la plus petite baisse de $\mathscr{H}$. En utilisant cette mesure d'agrégation, la nouvelle partition en $J-l$ classes maximise $\mathscr{H}$ parmi toutes les partitions en $J-l$ classes obtenues par agrégation de deux classes de la partition en $J-l+1$ classes.

3. Étape $l = J-1$: arrêt. La partition en une seule classe est obtenue.

La hauteur d'une classe $C = A \cup B$ dans le dendrogramme est définie par $h(C) = d(A,B)$. On a bien $h(C) \geq 0$ mais la propriété de croissance monotone de l'indice c'est-à-dire "$A \subset B \Rightarrow h(A) \leq h(B)$" n'a pas encore été démontrée. Notons qu'en pratique sur les jeux de données réelles ou simulées que nous avons utilisés, nous n'avons jamais observé de phénomènes d'inversion.

L'avantage de cette approche est qu'il n'est pas nécessaire de fixer au départ un nombre de classes de variables. L'approche fournit l'arbre hiérarchique associé qui permet de visualiser les associations successives entre les variables et aide au choix du nombre de classes de variables. Cependant l'approche peut se révéler coûteuse si le nombre de variables est important. Il est préférable dans ce cas de démarrer par une approche de type k-means (décrite dans Chavent et al. (2012a)). L'idée est d'effectuer une première réduction du nombre de variables par classification k-means puis d'effectuer une classification ascendante hiérarchique sur les variables synthétiques des classes obtenues. Cette technique est actuellement disponible dans le package ClusOfVar. Le calcul des coefficients initiaux, permettant d'écrire les variables synthétiques obtenues à partir des variables initiales, fera prochainement l'objet d'un ajout dans la nouvelle version de ce package.

2.8. Interprétation des variables synthétiques des classes de variables

Lorsque le choix du nombre de classes de variables est effectué, l'interprétation des variables synthétiques construites se révèle très importante dans la démarche méthodologique proposée.

Tout d'abord, la formule (3) des scores des composantes principales issues de PCAMIX donne les coordonnées des observations sur la variable synthétique de la classe C_k. Les informations portées par les variables de la classe sont résumées par cette variable synthétique. Nous utilisons alors la matrice $\mathbf{A}_k$ définie dans (5) qui joue un rôle fondamental dans l'interprétation et l'étiquetage des variables synthétiques des classes. Elle permet en effet d'obtenir le même type de règles d'interprétation qu'en ACP ou ACM. La seule différence est que les matrices $\mathbf{A}_k^1$, $\mathbf{D}^{-1/2}\mathbf{A}_k^2$ et $\mathbf{C}_k$ sont définies à l'intérieur d'une classe. Elles sont donc de dimension plus faible (dans le sens où seules les variables de la classe k sont prises en compte, et non pas l'ensemble des variables disponibles), ce qui implique une lecture simplifiée. De plus, la variable synthétique d'une classe étant la première composante principale issue de PCAMIX appliquée à la classe, nous allons nous intéresser seulement à la première colonne de $\mathbf{A}_k^1$, $\mathbf{D}^{-1/2}\mathbf{A}_k^2$ et $\mathbf{C}_k$.

Un premier élément intéressant est d'identifier les variables qui sont le plus liées à chacune des variables synthétiques. Pour les variables quantitatives, la première colonne de la matrice $\mathbf{A}_k^1$ permet de lire les corrélations des variables quantitatives à la variable synthétique de la classe (une sortie classique de l'ACP). Pour les variables qualitatives, la première colonne de la matrice $\mathbf{C}_k$ fournit les rapports de corrélation des variables de la classe à leur variable synthétique. Par ailleurs, lorsqu'on est en présence d'un ensemble de variables mixtes, la matrice $\mathbf{C}_k$ trouve tout son sens car elle permet de visualiser sur une échelle commune $[0,1]$ les liaisons entre les variables quantitatives et qualitatives d'une part et la variable synthétique d'autre part.

 ISSN: 2102-6238

Pour les variables qualitatives, l'interprétation fine des valeurs de la variable synthétique semble moins évidente au premier abord. En effet, on construit des classes de variables qualitatives alors que la variable synthétique construite dans chaque classe est quantitative. Dans la démarche méthodologique que nous proposons, il s'agit de lire cette variable synthétique comme une sorte de gradient. L'idée est de positionner les coordonnées des modalités des variables qualitatives de la classe sur cette variable synthétique (première colonne de la matrice $\mathbf{D}^{-1/2}\mathbf{A}_k^2$) afin de voir des regroupements de modalités apparaître. Plus précisément la variable synthétique prend un nombre limité de valeurs qui correspondent à des associations de modalités (si les variables qualitatives initiales sont liées). Cette interprétation est facilitée par le fait que la variable synthétique fait seulement référence aux variables de la classe et donc seules les modalités de ces variables ont des coordonnées sur cette variable synthétique (contrairement à l'ACM où, pour chaque composante, on visualise les modalités de l'ensemble des variables). Notons à ce titre que lorsque les variables sont mixtes, la propriété quasi-barycentrique de l'ACM est conservée : les coordonnées des modalités des variables sont les moyennes des scores des composantes principales standardisées sur les observations possédant cette modalité. Cette lecture de la variable synthétique comme un gradient sera illustrée dans la section 3.

3. Application à la prise en compte de l'environnement par les agriculteurs

3.1. Description des données et de la problématique

En 2005, une enquête postale a été menée à l'échelle nationale auprès d'agriculteurs, par une équipe de sociologues d'Irstea (ex-Cemagref) de Bordeaux. Cette étude se situe dans un contexte de transformation de la politique agricole. En effet, d'une agriculture qui nourrit les hommes, l'agriculture est également reconnue depuis une trentaine d'années pour de nouvelles finalités comme la protection des ressources naturelles ou encore la vitalité des espaces ruraux. C'est dans ce contexte de multifonctionnalité de l'agriculture, concept apparu en 1992 au sommet de Rio, que l'enquête porte sur "la prise en compte de l'environnement par les agriculteurs". Mais cette interrogation ne peut se résumer à une simple variable binaire. Il s'agit d'une notion plus complexe, qui évolue en fonction des normes environnementales ou sanitaires, des cahiers des charges des aides publiques, des préoccupations environnementales du monde agricole et non-agricole, etc. En conséquence, s'orienter vers une agriculture environnementale peut être perçu comme un "changement à la fois technique, cognitif et structurel" (Candau et al. (2005)). Les agriculteurs peuvent aussi être confrontés à des enjeux qui remettent en cause leur activité plus fondamentalement que ne le fait la protection de l'environnement. C'est pourquoi l'étude aborde cette problématique à l'aide d'un questionnaire structuré en quatre grandes parties. Elle interroge à la fois les transformations actuelles du métier d'agriculteur, les pratiques en faveur de l'environnement, la conception de l'environnement et de la nature par les agriculteurs ainsi que leur évaluation des mesures agro-environnementales (MAE [1]). Plus précisément, la conception du métier est appréhendée au travers de questions générales sur leur profession (q1_1 à q1_4) ou par le biais de variables plus spécifiques, qui portent sur les attraits de l'activité (q5_1 à q5_6), les finalités poursuivies (q6_1 à q6_7) ou encore les difficultés rencontrées dans l'exercice de leur

[1] MAE : dans le questionnaire, il s'agit de toute mesure réglementaire ou incitative visant la protection de l'environnement

 ISSN: 2102-6238

activité (q2_1 à q2_7 et q3_1 à q3_6). Les variables qui ont trait à l'environnement s'intéressent plus précisément aux problèmes de l'environnement et à l'évaluation de leur gravité (q8_1 à q8_2, q9_1 à q9_6), à la relation agriculture-environnement dans les 20 prochaines années (q10_1 à q10_4) et au rapport que les agriculteurs entretiennent avec la nature (q12_1 à q12_5) en les interrogeant en particulier sur les mesures agro-environnementales, leur évaluation et les difficultés de mise en œuvre (q13_1 à q13_5, q15_1 à q15_6, q18_1 à q18_9). Le jeu de données est ainsi constitué de 67 variables qualitatives, à deux ou trois modalités, observées sur un échantillon de 544 individus. La liste détaillée des variables est présentée dans l'annexe 2.

Les données concernent donc un ensemble de variables relatives à des thématiques diverses. Dans ce cas, comment répondre à la question complexe de la "prise en compte de l'environnement par les agriculteurs" ? Une première solution passe par une réduction de l'information. Les méthodes d'analyse des données se prêtent parfaitement au traitement de ce type d'enquête, en permettant l'étude simultanée de plusieurs variables. Pour répondre à cette problématique, une stratégie habituelle consiste à réaliser une analyse factorielle des données (ici une ACM vu la nature exclusivement qualitative des données). Nous optons pour une approche différente par classification de variables, avec la méthode ClustOfVar décrite dans la section 2.

3.2. La classification de variables : de 67 variables initiales à une partition en neuf classes de variables

L'originalité de notre approche par rapport à une démarche "classique" d'analyse de données d'enquêtes réside dans le remplacement de l'étape d'ACM par un algorithme de classification de variables. Cette approche par classification de variables permet de construire simultanément aux classes de variables les variables synthétiques associées. Dans notre étude de cas, nous verrons que ces variables synthétiques sont faciles à étiqueter. L'idée de la démarche est de dégager grâce à la construction de variables synthétiques des premières pistes d'interprétation quant aux rapports différenciés des agriculteurs à l'environnement.

Choix du nombre de classes de variables. Le package R intitulé ClustOfVar (Chavent et al. (2012a)), disponible sur le CRAN, est utilisé pour réaliser la classification ascendante hiérarchique des variables. Le dendrogramme issu de cette classification (voir la Figure 1) permet d'examiner les agrégations successives de l'ensemble des variables et de visualiser les liaisons qui existent entre elles. L'observation de l'arbre est complétée par la Figure 2 représentant l'évolution du critère d'agrégation en fonction du nombre de classes (et utilisé pour indicer la hiérarchie). À chaque étape de la classification ascendante, ce critère mesure la perte en homogénéité lorsque deux classes sont agrégées. Un coude dans l'évolution de ce critère correspond à l'agrégation de classes très différentes. Cependant il est délicat de choisir un nombre de classes de variables avec ces deux graphiques. Il n'est pas évident de détecter un saut dans l'arbre hiérarchique ou, de façon équivalente, d'identifier une "cassure" nette dans la Figure 2. Il semble toutefois qu'un nombre de dix classes environ soit pertinent d'un point de vue statistique.

Cependant le choix du nombre de variables synthétiques ne se résume pas à des arguments statistiques. Il est également dicté par l'interprétation que l'on peut faire de ces regroupements de variables. Dans notre étude de cas, l'accent est mis sur la compréhension des classes de variables et leur lecture en lien avec la problématique posée. En regardant l'interprétation des

 ISSN: 2102-6238

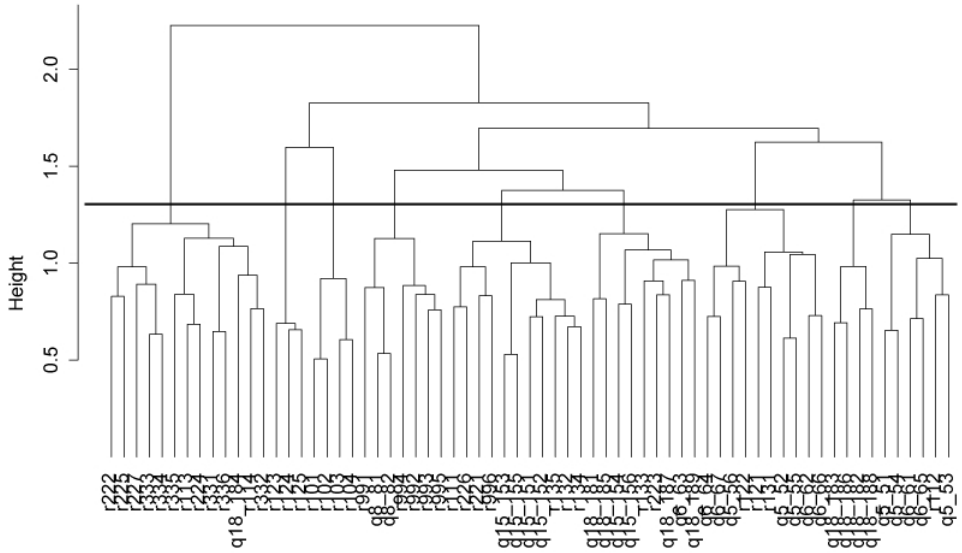

FIGURE 1: Dendrogramme issu de la classification ascendante hiérarchique des 67 variables qualitatives (coupure en 10 classes indiquée par le trait horizontal)

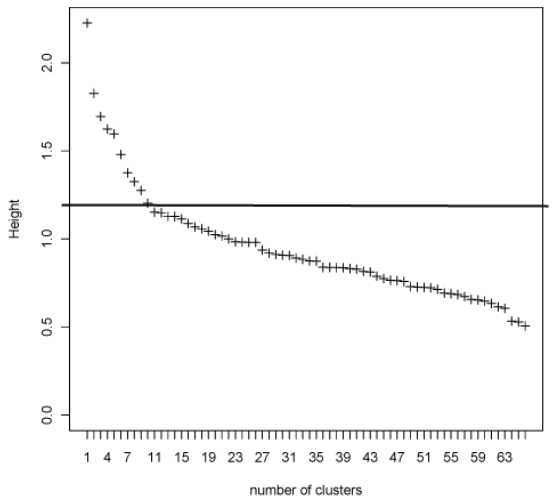

FIGURE 2: Évolution du critère de classification des 67 variables qualitatives

dix classes, nous constatons que deux d'entre elles sont composées de variables qui sont proches dans leur thématique. C'est pourquoi la partition en neuf classes, qui réunit à l'étape suivante de classification ces deux groupes de variables, est finalement retenue. Le regroupement des variables dans ces neuf classes est pertinent et riche en information pour les sociologues.

Concernant le choix du nombre de classes de variables, une dernière remarque concerne la structuration du questionnaire. Comme mentionné précédemment, la construction du questionnaire a été organisée autour de quatre grandes parties. L'objectif de la classification de variables est

 ISSN: 2102-6238

de regrouper les variables fortement liées entre elles, c'est-à-dire celles sur lesquelles il y a un lien dans la façon dont les individus ont répondu aux questions. Il est intéressant de remarquer à ce titre que le nombre de quatre classes ne semble pas ressortir. D'autre part, la partition des variables en quatre groupes issue de la classification des variables ne reprend pas ces parties du questionnaire. Ce premier résultat sur la classification des variables montre que les individus se ressemblent sur les choix qu'ils ont effectués sur les neuf groupes de variables et non uniquement au travers de ces quatre grands groupes de questions, ce qui constitue en soi un premier résultat intéressant pour l'interprétation sociologique.

Composition des neuf classes de variables. Les neuf classes de variables sont de tailles différentes, allant de trois variables pour la classe 8 à 13 variables pour la classe 3 (voir le Tableau 1). Le rapport de corrélation entre chaque variable qualitative et le représentant synthétique quantitatif de la classe (indiqué entre parenthèses) montre que les classes d'effectif plus faible ont des variables qui sont plus fortement reliées à la variable synthétique. On peut avancer comme explication que, pour les plus grandes classes, certaines valeurs sont plus faibles car elles regroupent des variables de thématiques plus diversifiées.

TABLE 1. *Partition des 67 variables qualitatives en neuf variables synthétiques, homogénéité des classes et pourcentage d'inertie expliquée par leur variable synthétique*

Classe	**1**	**2**	**3**	**4**	**5**
Nombre de variables	*11*	*6*	*13*	*9*	*10*
Variables (Rapport de corrélation)	q15_3 (0,39) q13_2 (0,38) q15_5 (0,35) q15_2 (0,34) q13_4 (0,30) q13_5 (0,31) q15_1 (0,26) q1_1 (0,19) q2_1 (0,13) q2_6 (0,13) q9_6 (0,05)	q5_1 (0,46) q6_1 (0,38) q5_3 (0,33) q5_4 (0,21) q6_5 (0,13) q1_2 (0,11)	q1_3 (0,28) q3_4 (0,26) q2_4 (0,26) q3_1 (0,23) q3_6 (0,22) q3_2 (0,21) q3_3 (0,18) q2_5 (0,18) q2_7 (0,18) q2_2 (0,14) q1_4 (0,12) q3_5 (0,12) q18_4 (0,01)	q2_3 (0,29) q2_3 (0,28) q2_3 (0,21) q2_3 (0,18) q2_3 (0,17) q2_3 (0,13) q18_8 (0,13) q6_3 (0,11) q18_1 (0,01)	q5_5 (0,47) q6_6 (0,40) q5_2 (0,35) q6_2 (0,21) q13_1 (0,11) q6_4 (0,10) q5_6 (0,06) q6_7 (0,05) q12_1 (0,03) q12_2 (0,01)
Homogénéité de la classe	*2,84*	*1,62*	*2,38*	*1,50*	*1,79*
Pourcentage d'inertie expliquée	*18,91*	*27,06*	*18,27*	*13,67*	*17,90*

Classe	**6**	**7**	**8**	**9**
Nombre de variables	*7*	*4*	*3*	*4*
Variables (Rapport de corrélation)	q8_1 (0,48) q8_2 (0,41) q9_2 (0,28) q9_5 (0,25) q9_3 (0,20) q9_1 (0,19) q9_4 (0,18)	q10_1 (0,54) q10_2 (0,54) q10_4 (0,52) q10_3 (0,37)	q12_4 (0,58) q12_5 (0,54) q12_3 (0,53)	q18_3 (0,48) q18_6 (0,39) q18_8 (0,37)) q18_23 (0,32)
Homogénéité de la classe	*1,98*	*1,97*	*1,65*	*1,56*
Pourcentage d'inertie expliquée	*22,04*	*24,61*	*55,16*	*39,05*

L'homogénéité de chacune des classes est également précisée dans ce tableau. D'après la sous-section 2.4, elle est définie comme la plus grande valeur propre de l'ACM de la classe, c'est-

 ISSN: 2102-6238

à-dire la variance de sa variable synthétique. Pour comparer les valeurs les unes par rapport aux autres, nous calculons le pourcentage d'inertie de la classe expliquée par la variable synthétique. Pour cela, nous divisons l'homogénéité de la classe par la variance totale de la classe qui, nous le rappelons, vaut $\frac{m_k}{J_k} - 1$ où J_k désigne le nombre de variables de la classe C_k et m_k désigne le nombre de modalités des variables de C_k. Les variables synthétiques des classes 8 et 9 sont celles qui expliquent le plus grand pourcentage de variance, ce sont des classes avec peu de variables, relatives à des thématiques très proches. Par ailleurs, la classe 3 présente un pourcentage d'inertie nettement plus faible mais ceci est à mettre en balance avec le fait qu'il s'agit de la classe regroupant le plus de variables. Finalement, nous verrons plus loin que la classe 4 ayant le pourcentage le plus faible, regroupe des variables relatives à différentes thématiques.

Interprétation et étiquetage des neuf variables synthétiques. Comme annoncé dans la sous-section 2.6, la matrice $\mathbf{D}^{-1/2}\mathbf{A}_k^2$ issue de PCAMIX, contenant les coordonnées des modalités des variables qualitatives, constitue un point essentiel dans notre démarche méthodologique. En regardant les coordonnées des modalités sur les variables synthétiques (première colonne de cette matrice), on peut visualiser chaque variable comme une sorte de gradient. La Figure 3 présente sur une même échelle les regroupements de modalités associées à chaque variable synthétique. On voit que les valeurs négatives ou positives de chaque variable synthétique (quantitative) sont associées à un regroupement distinct de modalités des variables de la classe correspondante. Prenons pour exemple la lecture de la variable synthétique de la classe 3 (VS3). Il s'agit de la variable quantitative à laquelle les variables relatives aux difficultés rencontrées par les agriculteurs dans l'exercice de leur activité (q1_3, q3_4, q2_4, q3_1, q3_6, q3_2 ; voir tableau 1) sont le plus liées. Les valeurs positives de VS3 sont associées aux modalités "Pas d'accord" relatives aux questions sur des difficultés telles la main d'œuvre, la paperasserie, la mise aux normes, le prix des terres et le temps de travail. Les valeurs légèrement négatives et proches de zéro de VS3 sont associées aux modalités "D'accord" de ces mêmes questions. Ainsi les avis émis se partagent en deux pôles opposés : les agriculteurs qui considèrent que les difficultés sont importantes et relèvent de changements structurels et à l'inverse, ceux pour qui l'exercice de l'activité ne pose pas de problème. Sur ce principe, il est alors possible d'étiqueter chacune des variables synthétiques. Le Tableau 2 présente une version synthétique de cette lecture des variables synthétiques. Nous insistons sur le fait que l'étiquetage des variables synthétiques et au-delà la présentation des résultats sont simplifiés ici dans cet article méthodologique. Sur le plan purement sociologique, l'interprétation est bien évidemment plus fine et complexe.

3.3. De la classification de variables à la typologie d'individus

Une partition des individus en sept classes. L'approche par classification de variables a permis d'identifier neuf variables synthétiques, qui mettent en lumière des premiers éléments sur les caractéristiques distinguant les individus quant aux transformations de l'activité agricole (situation de l'environnement, lien avec le monde non-agricole, confiance en l'avenir, etc.). Pour compléter cette analyse de la prise en compte de l'environnement par les agriculteurs, nous réalisons une typologie des individus afin de mettre en évidence des profils types distincts. Plus précisément, nous effectuons une classification ascendante hiérarchique (CAH avec critère de Ward) sur les scores des individus mesurés sur les neuf variables synthétiques. Une étape de consolidation à

 ISSN: 2102-6238

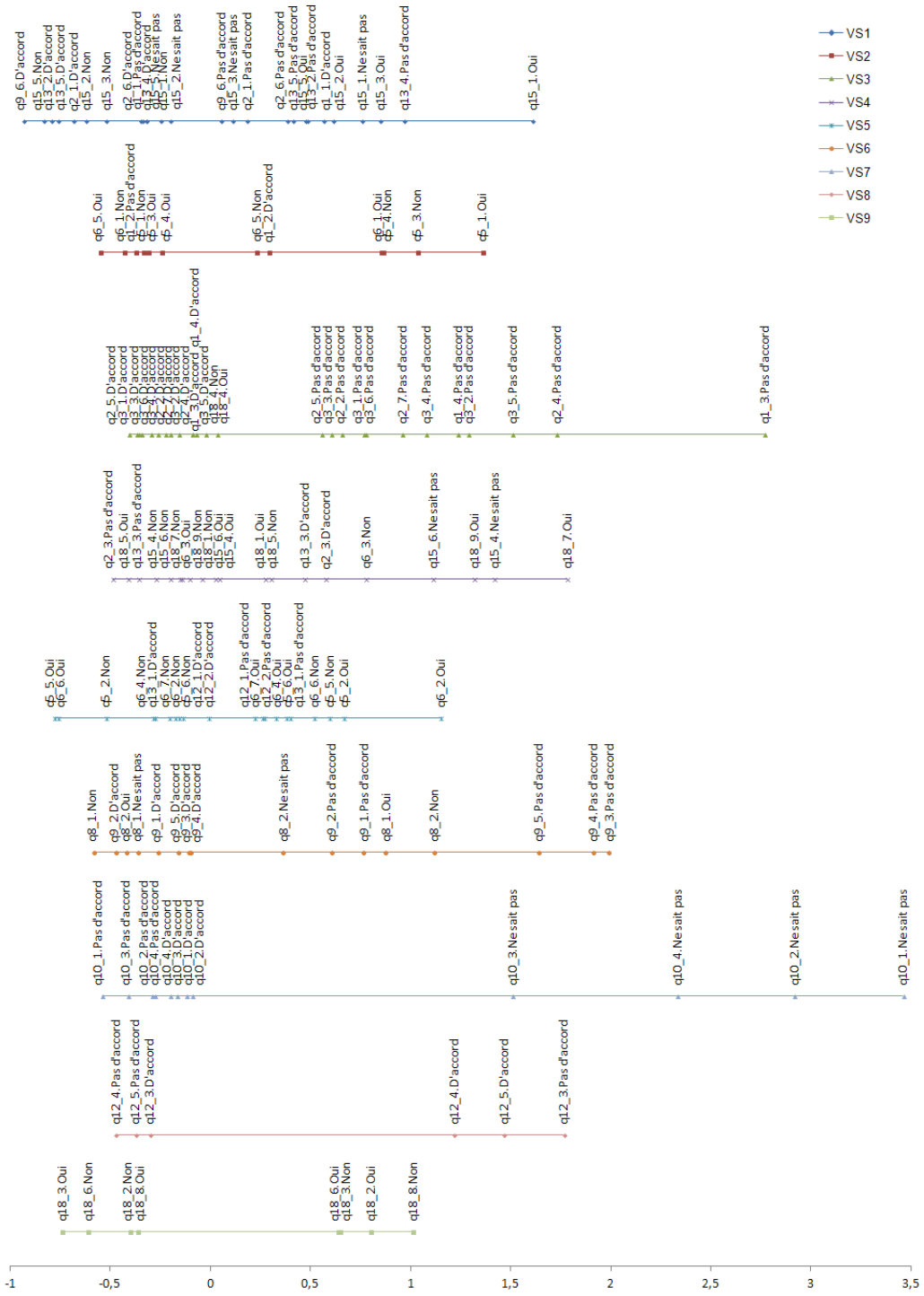

FIGURE 3: Gradients des neuf variables synthétiques

 ISSN: 2102-6238

TABLE 2. *Résumé des informations des neuf variables synthétiques*

	Label	Valeurs négatives	Valeurs positives
VS1	Relation avec le monde non-agricole	Lien difficile avec le monde non-agricole, MAE semblent être un frein à l'activité, problèmes d'environnement ignorés	Mesures environnementales bénéfiques pour l'activité et le lien avec le monde non-agricole
VS2	Attraits du métier	Indépendance, contact avec la nature, nourrir les hommes	Adaptation au marché, technique de pointe, activité motivante
VS3	Difficultés du métier, de son exercice	Difficultés nombreuses, de plusieurs ordres	Confiance en l'avenir, pas de difficulté
VS4	Adaptation du métier aux mesures environnementales et aspect économique du métier	Préoccupations économiques pour l'application des MAE et la finalité du métier	Difficulté d'adaptation du métier aux mesures en faveur de l'environnement et ses applications. Les mesures en faveur de l'environnement véhiculent une image ancienne de l'agriculture, incitent à revenir à des savoir-faire anciens
VS5	Finalité du métier	Adaptation, évolution	Protection, histoire familiale, patrimoine
VS6	Situation de l'environnement	Inquiétude, attention portée à la situation environnementale	Pas d'inquiétude, rejet de la situation environnementale
VS7	Relation agriculture-environnement dans 20 ans	Avis tranché (d'accord ou pas)	Indécis
VS8	Zones peu productives	Entretien de ces zones	Pas d'entretien de ces zones
VS9	MAE	Difficultés d'ordre administratif	Difficultés d'ordre économique, de travail

l'aide de l'algorithme des k-means est également réalisée afin de stabiliser la typologie obtenue.

L'examen du dendrogramme et de l'histogramme des indices de niveau de la CAH appliquée à nos données, révèle un saut pour trois ou sept classes d'individus. Par rapport à la problématique étudiée dans l'analyse sociologique, une typologie en trois classes ne présente pas un grand intérêt. La typologie en sept classes est plus pertinente, notamment pour rendre compte de la diversité des profils qui se sont dessinés avec l'étape préalable de classification de variables. Le Tableau 3 présente les effectifs des classes d'agriculteurs. Deux classes sont de taille très petite (la classe 3 et la classe 7 qui regroupent chacune environ 6% des individus). Cependant après une analyse fine de ces classes, nous observons dans le dendrogramme qu'elle se forment tôt et s'agrègent très tardivement aux autres classes. À ce titre, elles méritent d'être retenues car elles contiennent des agriculteurs aux caractéristiques bien particulières. Trois classes sont de taille relativement identique (la classe 1 avec 21% des individus ainsi que les classes 4 et 5 avec près de 20% chacune) et regroupent près des 2/3 de l'échantillon.

TABLE 3. *Composition des sept classes d'individus*

Classe d'individus	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Effectif	115	85	34	103	108	69	30
Pourcentage	21,1	15,6	6,3	18,9	19,9	12,7	5,5

Visualisation de la partition des agriculteurs. Nous souhaitons analyser la qualité de la partition des individus que nous avons construite. Pour cela, il est possible de réaliser des

 ISSN: 2102-6238

graphiques qui sont plutôt “classiques” lorsqu’on fait de la classification d’individus, par exemple la projection des individus habillés en fonction de leur classe d’appartenance. Ce type de graphique nécessite dans notre cas une étape préalable. En effet, l’algorithme de classification de variables que nous utilisons n’impose pas de contraintes d’orthogonalité entre les variables synthétiques. Il est donc important d’analyser si les variables synthétiques sont corrélées ou non. Le tableau 4 montre que ces variables sont faiblement liées et apportent donc des informations bien distinctes. Seules les variables synthétiques 1 et 6 sont corrélées négativement (-0,33). Un test de corrélation de Pearson confirme la significativité de cette corrélation avec une p-valeur égale à 10^{-5}. Au vu de l’interprétation des VS faite précédemment, ce résultat semble intuitif. Les agriculteurs qui portent une attention particulière à la situation environnementale ont tendance à penser que les MAE sont bénéfiques pour l’activité et le lien avec le monde non-agricole (sans pouvoir visualiser ici une relation de cause à effet).

TABLE 4. *Corrélations entre les variables synthétiques des classes*

Classe	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	1	0,11	0,14	0,08	-0,11	**-0,33**	-0,03	-0,02	0,18
2		1	0,06	-0,01	-0,12	0,04	-0,12	-0,03	0,15
3			1	-0,01	-0,01	0,09	0,06	0,05	0,05
4				1	0,08	-0,09	0,01	-0,01	0,05
5					1	0,04	0,05	0,01	0,04
6						1	-0,02	-0,03	-0,12
7							1	-0,08	-0,03
8								1	0,07
9									1

Nous réalisons donc une ACP normée sur les neuf variables synthétiques. En conservant tous les axes, la totalité de l’information est conservée. L’intérêt de cette étape est de pouvoir rigoureusement réaliser et visualiser des projections dans des plans orthogonaux. Notons que dans notre application, la première et la seconde composantes principales issues de l’ACP correspondent respectivement aux variables synthétiques VS1 et VS2, mais il s’agit d’un pur hasard car le nom des variables synthétiques ne traduit pas un quelconque ordre d’importance. La Figure 4 présente ainsi la projection des individus habillés selon leur appartenance aux sept classes dans ce premier plan factoriel. Seule la projection dans ce plan est présentée pour éviter de surcharger l’article. Ce graphique montre des classes qui sont relativement bien homogènes et séparées les unes des autres, avec une qualité de projection des individus satisfaisante. Par exemple, les classes 1 et 2 s’opposent nettement à la classe 5 sur la VS1. D’après le Tableau 2, cette VS caractérise les relations avec le monde non agricole. Ainsi les individus de la classe 5, ayant des valeurs négatives sur cette VS, ont un lien difficile avec ce monde et pensent que les MAE sont un frein à leur activité. Au contraire les agriculteurs des classes 1 et 2 n’expriment pas de difficulté particulière quant à leur lien avec le monde non agricole.

 ISSN: 2102-6238

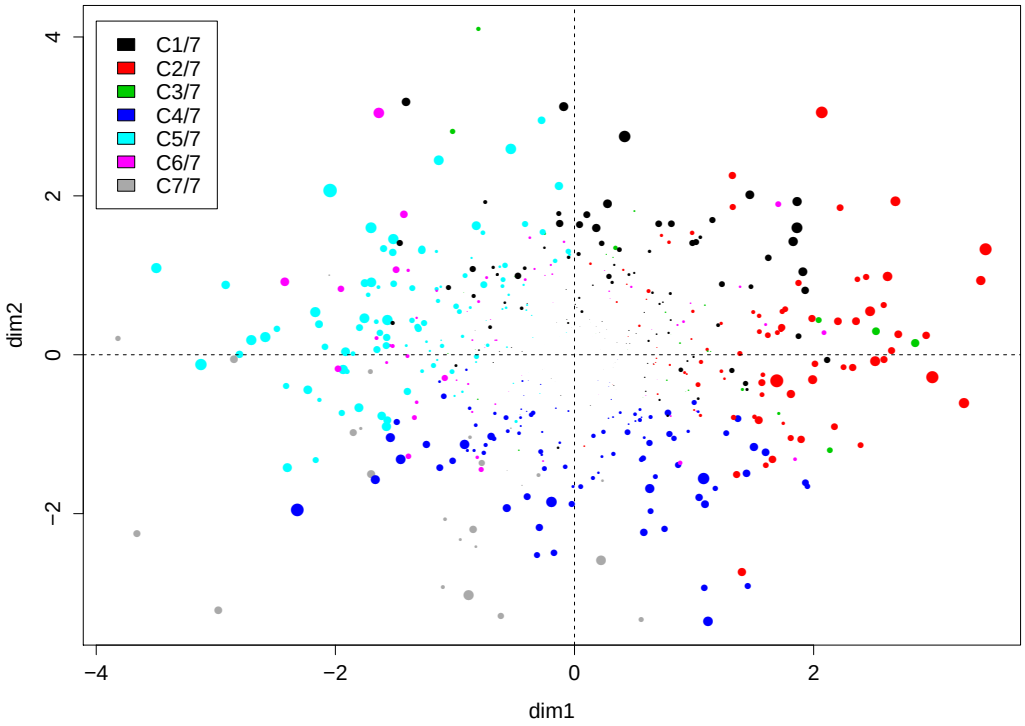

FIGURE 4: Nuage des individus habillés en fonction de leur classe (*dim1 et dim2 sont les deux premiers axes issus de l'ACP normée réalisée sur les neuf variables synthétiques, ils correspondent dans ce cas à VS1 et VS2. Cette étape supplémentaire est nécessaire pour une projection dans un plan orthogonal.)*

Critères internes de validité de la partition. Bien qu'il soit délicat de dire qu'une partition est meilleure qu'une autre en classification (non supervisée), il existe des indices internes de validité pouvant renseigner l'utilisateur sur la qualité de la partition (voir Gordon (1999) ou Mirkin (2005)). Par exemple, la silhouette moyenne de la partition est une mesure usuelle qui évalue la correspondance entre une structure de classification et les données à partir desquelles elle a été générée. Il s'agit d'un degré de confiance dans l'affectation des observations aux classes. Des mesures de séparation telle que la plus petite distance entre un point de la classe et un point appartenant à une autre classe peuvent également être calculées. Les calculs de ces critères, recensés dans le Tableau 5, permettent de valider la partition d'un point de vue statistique, puisqu'ils montrent bien que les classes sont homogènes entre elles et hétérogènes les unes par rapport aux autres.

Afin de tester la stabilité de la partition des individus, nous utilisons la méthode du rééchantillonnage par sous-échantillons. Plus précisément, nous générons à partir de l'échantillon initial de taille n, B copies en sélectionnant aléatoirement et sans remise $\alpha \times I$ observations, avec $0 < \alpha < 1$. Il s'agit alors de comparer les partitions obtenues en sept classes sur les sous-échantillons à

 ISSN: 2102-6238

TABLE 5. *Critères internes de validité de la partition*

Critères internes de validité	**Classe d'individus**						
	1	2	3	4	5	6	7
Maximum des distances entre classes	0,56	0,59	0,79	0,56	0,61	0,59	0,81
Plus petite distance entre un point de la classe et un point appartenant à une autre classe	0,07	0,08	0,16	0,07	0,09	0,11	0,13
Distance moyenne entre un point d'une classe et les points des autres classes	0,38	0,41	0,48	0,48	0,39	0,41	0,51
Silhouette moyenne de la partition	0,11						

la partition originale en sept classes (obtenue sur les données initiales). Cette comparaison est seulement faite sur les mêmes $\alpha \times I$ individus de la typologie initiale (et complète). Pour cela, nous utilisons le critère de Rand non ajusté (introduit par Rand (1971)), qui permet de mesurer le degré de similarité entre deux partitions ayant un même nombre de classes et effectuées sur les mêmes individus. Il évalue le nombre de paires de points qui sont classées ensemble dans les deux partitions et le nombre de paires de points qui sont dans deux classes différentes dans les deux partitions. Il s'interprète donc en termes de pourcentage d'accord entre les deux partitions et prend ses valeurs entre 0 (aucune concordance) et 1 (parfaite concordance). L'agrégation va consister dans notre cas à calculer l'indice de Rand moyen sur l'ensemble des partitions obtenues sur les sous-échantillons. L'idée sous-jacente est que si la partition est stable, de légères modifications du jeu de données ne doivent pas fondamentalement modifier les résultats de la classification (structure cachée forte). Nous appliquons cette technique avec $\alpha = 0,8$ et $B = 100$. Le critère de Rand moyen obtenu est égal à 0,70. Il traduit une partition stable, et que l'on peut alors considérer comme valide d'un point de vue statistique.

Une étape explicative de discrimination. Une autre façon de valider cette partition est d'utiiser une étape explicative de discrimination (méthode CART), qui permet de préciser les règles de construction des classes. La méthode CART est appliquée en utilisant les variables ayant permis de construire les classes (les neuf variables synthétiques quantitatives) comme variables explicatives. Nous observons que le nombre de règles avec les neuf variables synthétiques issues de ClustOfVar est faible (Figure 5). Cela signifie que l'affectation des individus est simple et que les classes sont homogènes et distinctes. L'arbre de la segmentation après élagage comporte ainsi peu de feuilles. Nous vérifions que le faible nombre de variables discriminantes n'explique pas à lui seul ce résultat. Pour cela, nous appliquons CART avec les 67 variables initiales et nous constatons que le nombre de règles n'a pas augmenté, ce qui valide bien notre partition en sept classes d'agriculteurs. Par ailleurs, la lecture des règles d'affectation est plus facile avec les variables synthétiques comme variables discriminantes qu'avec les 67 variables qualitatives initiales. Notons que l'utilisation de ces variables comme variables discriminantes est rendue possible par le fait que, dans notre application, les variables synthétiques ont pu être analysées comme des gradients et sont finement étiquetées.

Interprétation des classes d'individus à l'aide des variables synthétiques. L'interprétation des classes d'agriculteurs nous permet de répondre à la problématique posée. Une partition s'avère

 ISSN: 2102-6238

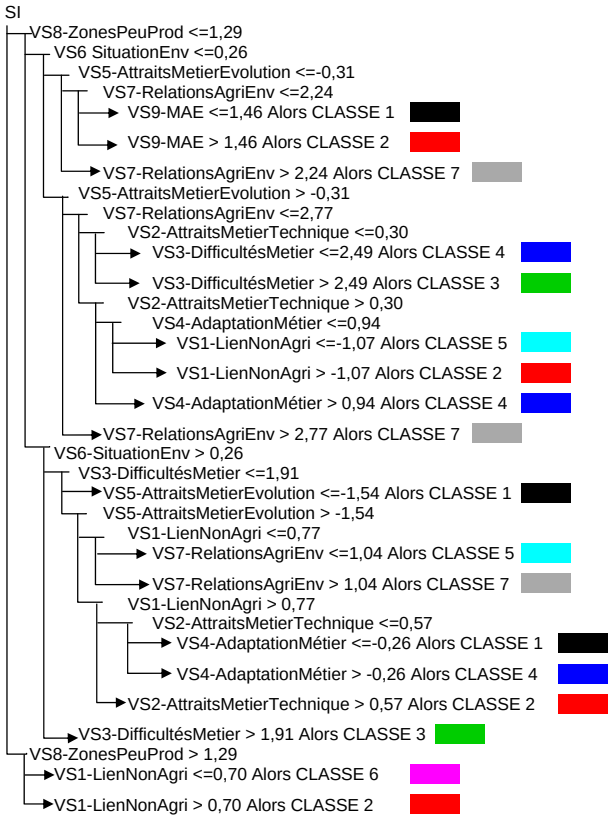

FIGURE 5: Arbre de décision pour la partition issue de l'approche par classification de variables

Lecture : Si la valeur de la variable synthétique 8, qui regroupe les variables relatives aux zones peu productives, est supérieure à 1,29 (pas d'entretien) et si la valeur de la variable synthétique 1, qui regroupe les variables relatives aux relations avec le monde agricole, est inférieure à 0,70 (lien difficile), alors l'individu se classe dans la classe 6.

intéressante lorsque l'on décrit les classes par les individus qui la composent et/ou les variables qui les caractérisent. Ici l'analyse par les individus ne présente pas d'intérêt en raison de leur anonymat. Par la réduction du nombre de variables via ClustOfVar, nous pouvons interpréter la partition des agriculteurs au regard des neuf variables synthétiques et non des 67 questions initiales. Le tableau 6 fournit la valeur moyenne de chaque variable synthétique dans les sept classes de la typologie.

La mise en parallèle avec l'étiquette des variables synthétiques du Tableau 2 constitue une aide à l'interprétation. Dans notre étude, la lecture des résultats est nettement facilitée. Plus précisément, nous comparons pour chaque variable synthétique la moyenne par classe d'individus à celle de l'échantillon total. Notons que cette moyenne est nulle car les variables synthétiques construites avec ClustOfVar sont centrées. Le Tableau 6 présente en gras respectivement les moyennes négatives (positives) qui sont significativement inférieures (supérieures) à 0 (p-valeur inférieure à 10^{-3}). Notons que ces tests n'ont pas de réelle valeur statistique car les variables synthétiques ont été utilisées pour créer les groupes d'individus. Ils servent donc seulement à

 ISSN: 2102-6238

TABLE 6. *Moyenne des variables synthétiques pour les sept classes d'individus. En gras : valeurs significativement inférieures ou supérieures à la moyenne de la VS dans l'échantillon total (égale à 0) ; p-valeur inférieure à* 10^{-3}.

Variable synthétique	Classe d'individus						
	1	2	3	4	5	6	7
1	0,517	**1,476**	0,668	0,418	**-1,503**	-0,886	-0,909
2	0,175	**1,176**	-0,160	**-0,695**	-0,022	-0,372	-0,499
3	-0,101	0,162	**4,026**	-0,523	-0,653	-0,322	0,251
4	-0,548	0,241	0,356	**0,823**	-0,278	-0,272	-0,185
5	**-1,471**	0,287	0,343	**0,826**	0,403	0,025	0,092
6	-0,289	**-0,747**	0,691	**-0,850**	**1,428**	0,039	0,130
7	-0,398	-0,200	-0,036	-0,022	-0,458	-0,313	**4,578**
8	-0,432	0,079	0,030	-0,513	-0,573	**2,437**	-0,381
9	-0,388	**1,349**	0,316	-0,416	-0,179	-0,079	-0,434

l'interprétation en indiquant les variables synthétiques qui caractérisent les classes d'individus. On voit que certaines classes (1, 3, 6, 7) peuvent être caractérisées par une seule variable synthétique, ce qui facilite leur interprétation. Certes, on retrouve ici les classes d'effectifs faibles (3 et 7), pour lesquelles on s'attend à une caractérisation concise. Cependant la classe 1, qui comptabilise le plus grand nombre d'individus (près d'un quart de l'échantillon), est également décrite par une seule variable synthétique. Ces résultats soulignent l'homogénéité à l'intérieur des groupes d'individus et l'hétérogénéité entre eux. Pour les trois classes d'individus restantes, la caractérisation est moins marquée, une interprétation au travers de plusieurs variables synthétiques est nécessaire. Cependant la compréhension de ces groupes reste relativement aisée au vu du faible nombre de variables synthétiques.

3.4. "Validation" de la typologie des agriculteurs

Une interprétation sociologique riche et pertinente. L'étiquetage des variables synthétiques est une aide efficace pour l'interprétation des classes d'individus. Nous allons voir ici que la typologie obtenue est pertinente et riche en information sur le volet sociologique.

Ainsi, la ***classe 1*** est caractérisée par une valeur moyenne négative de la VS5 ($\overline{VS5} = -1,471$). Les agriculteurs de cette classe sont intéressés par le changement, ils aiment leur métier parce que les personnes qui l'exercent doivent évoluer constamment et ils considèrent que les MAE leur demandent de maîtriser des techniques de pointe.

Les agriculteurs de la ***classe 2*** sont convaincus de la réalité des problèmes d'environnement et estiment qu'ils ne sont pas exagérés ($\overline{VS6} = -0,747$). Les mesures en faveur de l'environnement occupent une place importante dans le processus de production ($\overline{VS1} = 1,476$) et les démarches administratives liées aux MAE ne leur posent pas problème. Ils acceptent au contraire que l'activité agricole soit régulée par les pouvoirs publics et restent néanmoins attachés au marché ($\overline{VS2} = 1,176$) qui, selon eux, donne aussi des orientations à suivre. Cette catégorie d'agriculteurs éprouve conjointement des difficultés avec les dimensions entrepreneuriales de leur activité : ils dénoncent la charge de travail et les investissements nécessaires à la mise en œuvre des mesures en faveur de l'environnement ($\overline{VS9} = 1,349$).

Pour la ***classe 3*** on note une valeur moyenne positivement élevée de la VS3 ($\overline{VS3} = 4,026$), ce qui nous amène à considérer que cette classe est définie par les agriculteurs qui sont confiants en

 ISSN: 2102-6238

l'avenir et qui semblent exercer leur activité sans difficulté. Ils ne partagent pas bon nombre des difficultés proposées par l'enquête (paperasserie, prix des terres, main d'œuvre, etc.).

Enfin dans la ***classe 4***, on trouve des agriculteurs particulièrement attentifs à la protection de l'environnement ($\overline{VS6} = -0,850$) qu'ils considèrent difficile à concilier avec le progrès technique ($\overline{VS4} = 0,823$). Protéger les ressources naturelles et le paysage est, pour eux, une des premières finalités de leur activité ($\overline{VS5} = 0,826$). Si cette préoccupation environnementale laisse entrevoir des individus en questionnement et propices à remettre en cause certaines pratiques, l'évolution constante de leur activité ne les intéresse pas plus que cela ($\overline{VS2} = -0,695$). Ils déclarent en effet que le changement permanent n'est pas ce qui rend leur métier attrayant. Ils pensent même que les mesures environnementales réactivent des savoir-faire anciens.

Les agriculteurs de la ***classe 5*** rejettent les préoccupations environnementales, pensent que la gravité des problèmes d'environnement est exagérée et que la situation n'est pas inquiétante ($\overline{VS6} = 1,428$). Ils sont en même temps très critiques vis-à-vis des MAE qui sont pour eux un frein à l'activité et à leurs projets ($\overline{VS1} = -1,503$).

Les individus de la ***classe 6*** peuvent être jugés comme étant adeptes de la déprise agricole car ils ont une valeur positive pour la VS8 ($\overline{VS8} = 2,437$).

La ***classe 7*** est associée à une valeur moyenne positive de la VS7 ($\overline{VS7} = 4,578$). Autrement dit, ces agriculteurs ne se projettent dans aucun scénario de prospective proposé, sans les rejeter pour autant. L'avenir leur paraît incertain.

4. Discussion et conclusion

Une démarche originale pour répondre à la question complexe de la prise en compte de l'environnement par les agriculteurs. Dans cet article, nous proposons une démarche originale d'analyse des données dans la mesure où la classification de variables est utilisée comme alternative à la première étape classique d'analyse factorielle pour la typologie d'observations. L'approche ClustOfVar permet de construire des groupes de variables liées ainsi que les variables synthétiques quantitatives associées aux classes. Nous répondons ainsi à la question complexe de la prise en compte de l'environnement par les agriculteurs de façon "globale", en analysant les liaisons entre les variables ainsi que les ressemblances entre individus. En effet, la construction des variables synthétiques nous permet de distinguer quelles sont les "thématiques" (regroupements de variables) qui définissent des similitudes ou au contraire des différences au niveau des agriculteurs. La typologie des individus nous permet ensuite de préciser ces tendances et de dégager des profils-types d'individus.

Les variables synthétiques : souplesse dans la construction. La classification de variables permet de construire des variables synthétiques qui préservent au mieux les liaisons entre les variables initiales. En effet, contrairement aux méthodes d'analyse factorielle (ACP, ACM), leur construction ne sert pas un pourcentage d'inertie expliquée : identifier une première composante principale qui explique le plus grand pourcentage d'inertie initiale contenue dans le nuage de points, puis une seconde qui lui est orthogonale et qui explique un grand pourcentage d'inertie et ainsi de suite. Avec ces méthodes, on peut concevoir que des informations relatives à la structure des observations puissent être masquées par la création de ces composantes non corrélées qui visent seulement à reconstruire au mieux la variance initiale. La classification de variables supprime au

 ISSN: 2102-6238

contraire l'information redondante et la création des variables synthétiques se fait au vu de la réorganisation des variables en classes homogènes. De plus, avec cette approche, l'algorithme de classification n'impose pas de contraintes d'orthogonalité entre les variables synthétiques, ce qui offre plus de flexibilité dans la construction. Notons qu'une perspective à ce travail consisterait à proposer une méthode qui optimiserait simultanément le critère d'homogénéité de la classification de variables et le critère relatif à la typologie des observations. Ainsi on chercherait simultanément à construire des classes de variables liées tout en cherchant à identifier des profils-types d'individus. Un autre point intéressant dans le cas de données mixtes serait l'utilisation de poids différents pour équilibrer la part des deux types de variables quantitatives et qualitatives.

Nous avons vu que les variables synthétiques peuvent être lues comme une sorte de gradient. Sur nos données, leur interprétation est relativement aisée et nous permet de les étiqueter. Cette lecture simplifiée s'explique tout d'abord par le fait qu'avec la classification de variables, une variable synthétique résume l'information d'un sous-ensemble restreint de variables : les variables de la classe considérée. Ainsi seules les modalités des variables de la classe ont des coordonnées sur la variable synthétique. Sur ce jeu de données, l'utilisation d'une approche de classification de variables permet de simplifier l'interprétation des relations entre les variables d'une part et les modalités d'autre part.

Enfin l'apport des variables synthétiques pour construire et interpréter la classification des individus est incontestable.

Un apport en sociologie. L'apport de la démarche proposée réside tout d'abord dans le fait que les résultats obtenus permettent de mieux cerner la façon dont les agriculteurs conçoivent la protection de l'environnement en lien avec leur activité. Concernant les travaux en sociologie, l'utilisation d'une méthode d'analyse quantitative pour établir une typologie des agriculteurs sur leur prise en compte de l'environnement n'a pas encore été proposée dans la littérature française de ce domaine. Un article sur l'interprétation sociologique de la typologie obtenue ici est en cours de finalisation. Au-delà, même si l'application que nous décrivons repose sur des données qualitatives, nous insistons sur le fait que la méthodologie proposée est également applicable et valable pour des données quantitatives ou mixtes.

Références

Abdallah, H. and Saporta, G. (1998). Classification d'un ensemble de variables qualitatives. *Revue de Statistique Appliquée*, 46(4) :5–26.

Candau, J., Deuffic, P., Ginelli, L., Lewis, N., and Lyser, S. (2005). La prise en compte de l'environnement par les agriculteurs. Résultats d'enquête. Rapport d'étude, Cemagref.

Charrad, M. and Ben Ahmed, M. (2011). Simultaneous clustering : A survey. In *Pattern Recognition and Machine Intelligence*. Springer Berlin / Heidelberg.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2011). Classification de variables : le package clustofvar. In *43es Journées de Statistique (SFdS) , Tunis, TUN*.

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012a). Clustofvar : An r package for the clustering of variables. *Journal of Statistical Software*, 50(13) :1–16.

Chavent, M., Kuentz-Simonet, V., and Saracco, J. (2012b). Orthogonal rotation in PCAMIX. *Advances in Data Analysis and Classification*, 6(2) :131–146.

De Soete, G. and Carroll, J. (1994). K-means clustering in a low-dimensional Euclidean Space. In Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., and Burtschy, B., editors, *New Approaches in Classification and Data Analysis*, pages 212–219. Springer.

 ISSN: 2102-6238

De Soete, G. and Heiser, W. (1993). A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika*, 58 :545–565.

Derquenne, C. (2001). Classification de variables qualitatives : Une approche dynamique. In *34e Journées de Statistique, Nantes, FRA*.

DeSarbo, W., Jedidi, K., Cool, K., and Schendel, D. (1991). Simultaneous multidimensional unfolding and cluster analysis : An investigation of strategic groups. *Marketing Letters*, 2(2) :129–146.

Dhillon, I., Marcotte, E., and Roshan, U. (2003). Diametrical Clustering for Identifying Anticorrelated Gene Clusters. *Bioinformatics*, 19(13) :1612–1619.

Gordon, A. (1999). *Classification*. Chapman & Hall.

Govaert, G. and Nadif, M. (2009). Un modèle de mélange pour la classification croisée d'un tableau de données continues . In *CAP'09, 11e conférence sur l'apprentissage artificiel*.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., and Brown, P. (2000). 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2) :1–21.

Heiser, W. (1993). Information and classification. In Opitz, O., Lausen, B., and Klar, R., editors, *Information and Classification*, pages 162–173. Springer.

Kiers, H. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56(2) :197–212.

Lerman, I. (1990). Foundations of the likelihood linkage analysis classification method. *Applied Stochastics Models and Data Analysis*, 7(1) :63–76.

Lerman, I. (1993). Likelihood linkage analysis classification method : An example treated by hand. *Biochimie*, 75(5) :379–397.

Mirkin, B. (2005). *Clustering for Data Mining : A Data Recovery Approach*. Computer Science & Data Analysis. Chapman and Hall/CRC.

Plasse, M., Niang, N., Saporta, G., Villeminot, A., and Leblond, L. (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis*, 52 :596–613.

Qannari, E., Vigneau, E., and PH., C. (1998). Une nouvelle distance entre variables. Application en classification. *Revue de Statistique Appliquée*, 46(2) :21–32.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66 :846–850.

Soffritti, G. (1999). Hierarchical clustering of variables : a comparison among strategies of analysis. *Communications in Statistics - Simulation and Computation*, 28(4) :977–999.

Timmerman, M., Ceulemans, E., Kiers, H., and Vichi, M. (2010). Factorial and reduced K-means reconsidered. *Computational Statistics & Data Analysis*, 54(7) :1858–1871.

Vichi, M. and Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, 37(1) :49–64.

Vichi, M. and Saporta, G. (2009). Clustering and Disjoint Principal Component Analysis. *Computational Statistics & Data Analysis*, 53(8) :3194–3208.

Vigneau, E. and Qannari, E. (2003). Clustering of variables around latent components. *Communications in statistics Simulation and Computation*, 32(4) :1131–1150.

Vigneau, E., Qannari, E., Sahmer, K., and Ladiray, D. (2006). Classification de variables autour de composantes latentes. *Revue de Statistique Appliquée*, 54(1) :27–45.

Annexes

Annexe 1 : Illustration du calcul de la variable synthétique d'une classe

Pour illustrer le calcul de la variable synthétique d'une classe obtenue avec PCAMIX, nous utilisons le jeu de données "decathlon" présent dans ce package R. Les calculs sont réalisés sur une sous-matrice composée des observations 12 à 16 et des variables "100m", "javeline" et "competition". On suppose pour l'exemple que ces trois variables forment une classe, que l'on

 ISSN: 2102-6238

note k pour une harmonisation des notations avec la sous-section 2.5. Les variables quantitatives "100m" et "javeline" sont contenues dans la matrice $\mathbf{X}_k$ dont la version standardisée est donnée par $\tilde{\mathbf{X}}_k$. Notons que la standardisation s'obtient de façon classique par un centrage et une réduction en divisant par l'écart-type de la colonne.

$$\mathbf{X}_k = \begin{pmatrix} 11.33 & 57.44 \\ 11.36 & 54.68 \\ 10.85 & 70.52 \\ 10.44 & 69.71 \\ 10.50 & 55.54 \end{pmatrix} \text{ et } \tilde{\mathbf{X}}_k = \begin{pmatrix} 1.106 & -0.588 \\ 1.182 & -0.981 \\ -0.117 & 1.272 \\ -1.162 & 1.156 \\ -1.009 & -0.859 \end{pmatrix}.$$

La variable qualitative binaire "competition" est contenue dans $\mathbf{Z}_k$:

$$\mathbf{Z}_k = \begin{pmatrix} Decastar \\ Decastar \\ OlympicG \\ OlympicG \\ OlympicG \end{pmatrix}.$$

Dans le cas de variables qualitatives, le centrage-réduction s'effectue à partir du tableau disjonctif complet $\mathbf{G}$: $\tilde{\mathbf{Z}}_k = (\mathbf{I}_I - \mathbf{1}\mathbf{1}'/I)\mathbf{G}\mathbf{D}^{-1/2}$.

Sur ces données, on a : $\mathbf{G} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$; $\mathbf{D} = \begin{pmatrix} \frac{2}{5} & 0 \\ 0 & \frac{3}{5} \end{pmatrix}$ et $\mathbf{D}^{-1/2} = \begin{pmatrix} 1.581 & 0 \\ 0 & 1.291 \end{pmatrix}$.

On obtient donc $\tilde{\mathbf{Z}}_k = \begin{pmatrix} 0.949 & -0.775 \\ 0.949 & -0.775 \\ -0.632 & 0.516 \\ -0.632 & 0.516 \\ -0.632 & 0.516 \end{pmatrix}$.

La concaténation des matrices standardisées fournit la matrice $\mathbf{M}_k = \frac{1}{\sqrt{5}}(\tilde{\mathbf{X}}_k | \tilde{\mathbf{Z}}_k)$. La décomposition en valeurs singulières de $\mathbf{M}_k$ est donnée par $\mathbf{M}_k = \mathbf{U}_k \Lambda_k \mathbf{V}_k'$. Sur les données de l'illustration, on obtient la matrice $\mathbf{U}_k$ de dimension 5×3, la matrice Λ_k de dimension 3×3 et la matrice $\mathbf{V}_k$ de dimension 4×3 :

$$\mathbf{U}_k = \begin{pmatrix} -0.498 & 0.178 & 0.209 \\ -0.568 & 0.000 & -0.032 \\ 0.352 & 0.506 & -0.648 \\ 0.515 & 0.147 & 0.696 \\ 0.199 & -0.831 & -0.226 \end{pmatrix} ; \quad \Lambda_k = \begin{pmatrix} 1.547 & 0.000 & 0.000 \\ 0.000 & 0.749 & 0.000 \\ 0.000 & 0.000 & 0.217 \end{pmatrix}$$

$$\text{et} \quad \mathbf{V}_k = \begin{pmatrix} -0.596 & 0.481 & -0.643 \\ 0.498 & 0.850 & 0.173 \\ -0.487 & 0.168 & 0.578 \\ 0.398 & -0.137 & -0.472 \end{pmatrix}.$$

 ISSN: 2102-6238

Les scores des cinq individus sur les trois composantes principales sont définis dans la matrice $\sqrt{I}\mathbf{U}_k\Lambda_k$ de dimension 5×3 :

$$\begin{array}{c} \\ \begin{pmatrix} 1.723 & 0.297 & -0.101 \\ 1.964 & 0.001 & 0.016 \\ -1.217 & 0.847 & 0.314 \\ -1.783 & 0.247 & -0.337 \\ -0.688 & -1.392 & 0.109 \end{pmatrix} \end{array} \begin{array}{c} \textit{dim1} \quad \textit{dim2} \quad \textit{dim3} \\ \textit{Nool} \\ \textit{Bourguignon} \\ \textit{Sebrle} \\ \textit{Clay} \\ \textit{Karpov} \end{array}$$

La première colonne de cette matrice fournit la variable synthétique de la classe $\mathbf{y}_k = \sqrt{I}\lambda_k^1\mathbf{u}_k^1$ où $\mathbf{u}_k^1$ désigne la première colonne de $\mathbf{U}_k$. On lit les coordonnées des individus sur cette variable synthétique.

Pour interpréter les résultats, nous calculons la matrice $\mathbf{A}_k = \mathbf{V}_k\Lambda_k$ de dimension 4×3 :

$$\mathbf{A}_k = \begin{array}{c} \begin{pmatrix} 0.923 & 0.360 & 0.139 \\ -0.771 & 0.636 & -0.038 \\ -0.754 & 0.126 & 0.125 \\ 0.616 & -0.103 & -0.102 \end{pmatrix} \end{array} \begin{array}{c} \textit{dim1} \quad \textit{dim2} \quad \textit{dim3} \\ \textit{100m} \\ \textit{javeline} \\ \textit{Decastar} \\ \textit{OlympicG} \end{array}$$

Les deux premières lignes correspondent à la matrice $\mathbf{A}_k^1$ qui contient les corrélations des variables quantitatives de la classe aux trois composantes principales. Les deux dernières lignes forment la matrice $\mathbf{A}_k^2$, à partir de laquelle on peut calculer la matrice $\mathbf{DA}_k^2$. Cette dernière contient les coordonnées des modalités des deux variables qualitatives de la classe sur les composantes principales :

$$\mathbf{D}^{-1/2}\mathbf{A}_k^2 = \begin{array}{c} \begin{pmatrix} 1.192 & 0.199 & -0.198 \\ -0.795 & -0.133 & 0.132 \end{pmatrix} \end{array} \begin{array}{c} \textit{dim1} \quad \textit{dim2} \quad \textit{dim3} \\ \textit{Decastar} \\ \textit{OlympicG} \end{array}$$

À partir de $\mathbf{A}_k$ on obtient la matrice $\mathbf{C}_k$ qui fournit les corrélations au carré (resp. rapports de corrélation) entre les variables quantitatives (resp. qualitatives) et les composantes principales :

$$\mathbf{C}_k = \begin{array}{c} \begin{pmatrix} 0.851 & 0.130 & 0.019 \\ 0.594 & 0.405 & 0.001 \\ 0.947 & 0.026 & 0.026 \end{pmatrix} \end{array} \begin{array}{c} \textit{dim1} \quad \textit{dim2} \quad \textit{dim3} \\ \textit{100m} \\ \textit{javeline} \\ \textit{competition} \end{array}$$

Cette matrice permet de lire sur une même échelle [0,1] les liaisons entre les différents types de variables et les composantes principales. Ainsi la variable qualitative "competition" et la variable quantitative "100m" dont les valeurs sont proches de 1 sont très liées avec la variable synthétique $\mathbf{y}_k$ de la classe.

 ISSN: 2102-6238

Annexe 2 : Liste des 67 questions de l'enquête

Question	Variable	Libellé
1		**Trouvez-vous que ...**
	q1_1	Votre activité est perçue positivement par les non agriculteurs
	q1_2	Votre activité est motivante
	q1_3	Votre métier connaît de profonds changements
	q1_4	Vous êtes inquiet pour l'avenir de votre activité
2		**Votre métier en général, qu'est-ce qui vous semble difficile aujourd'hui ?**
	q2_1	L'augmentation de la fréquentation touristique
	q2_2	La baisse du nombre des agriculteurs
	q2_3	Les besoins en formation professionnelle
	q2_4	La paperasserie
	q2_5	La protection de l'environnement
	q2_6	Les relations avec les voisins non-agriculteurs
	q2_7	La transmission des exploitations
3		**Dans l'exercice de votre métier, que trouvez-vous difficile aujourd'hui ?**
	q3_1	La main d'œuvre
	q3_2	La mise aux normes
	q3_3	La nécessité de s'agrandir
	q3_4	Le prix des terres
	q3_5	Le prix et la commercialisation des produits
	q3_6	Le temps de travail
5		**Quels sont pour vous les principaux attraits de votre métier ?**
	q5_1	Être à la pointe de la technique
	q5_2	Être détenteur d'un patrimoine
	q5_3	Être en contact avec la nature
	q5_4	Être indépendant
	q5_5	Évoluer constamment
	q5_6	Perpétuer l'histoire familiale sur la région
6		**Aujourd'hui la finalité de votre métier c'est de ?**
	q6_1	Entreprendre en s'adaptant au marché
	q6_2	Entretenir les bâtiments anciens
	q6_3	Faire vivre votre famille
	q6_4	Maintenir et transmettre votre exploitation
	q6_5	Nourrir les hommes
	q6_6	Produire en s'adaptant aux attentes de la société
	q6_7	Protéger les ressources naturelles et le paysage
8		**Pensez-vous que ?**
	q8_1	On exagère la gravité des problèmes de l'environnement
	q8_2	La situation de l'environnement est inquiétante
9		**Selon vous les problèmes d'environnement sont l'affaire ...**
	q9_1	Des agriculteurs
	q9_2	Des associations de protection
	q9_3	De chaque consommateur
	q9_4	Des industriels
	q9_5	Des pouvoirs publics
	q9_6	De personne car ce ne sont pas des problèmes
10		**Les relations agriculture-environnement évoluent. Dans les 20 prochaines années, quels scenarii vous semblent les plus vraisemblables ?**

 ISSN: 2102-6238

	q10_1	L'agriculture sera plus liée à l'agroalimentaire et devra respecter des normes de qualité
	q10_2	L'environnement sera au cœur de l'agriculture avec des systèmes proches de l'agriculture biologique
	q10_3	L'Europe donnera le cadre général de la production et de l'environnement et la Région gèrera les objectifs plus précis
	q10_4	On aura d'un côté des zones intensives vouées à la production et de l'autre des zones vouées à la préservation
12		**Êtes-vous d'accord avec ces idées ?**
	q12_1	Je dois maîtriser la nature pour mon activité
	q12_2	Je dois m'adapter à la nature
	q12_3	Je dois entretenir les parties peu productives de mon exploitation pour qu'elles soient propres
	q12_4	J'évite d'intervenir sur les parties peu productives de mon exploitation pour laisser se développer la nature
	q12_5	Je n'interviens pas sur ces zones pour diminuer mon travail
13		**En tant qu'agriculteur, on vous invite (ou oblige) de plus en plus fréquemment à répondre à des mesures pour la protection de l'environnement. Pour votre activité, ces mesures ...**
	q13_1	Demandent de maîtriser des techniques de pointe
	q13_2	Empêchent de progresser
	q13_3	Incitent à revenir à des savoirs-faire anciens
	q13_4	Limitent votre liberté d'action
	q13_5	Touchent des domaines qui ne regardent que vous
15		**Estimez-vous que les mesures en faveur de l'environnement ?**
	q15_1	Incitent les jeunes à s'installer en agriculture
	q15_2	Permettent d'améliorer la qualité des produits
	q15_3	Renforcent la solidarité entre milieux agricole et non-agricoles
	q15_4	Sont un bon moyen pour limiter la production
	q15_5	Valorisent l'image de l'agriculture
	q15_6	Véhiculent une image ancienne de l'agriculture
18		**Dans l'application des mesures agri-environnementales, qu'est-ce qui vous semble le plus difficile ?**
	q18_1	Les changements techniques proposés
	q18_2	La charge de travail
	q18_3	Les contrôles
	q18_4	L'efficacité des mesures
	q18_5	Le faible montant de l'aide
	q18_6	L'investissement financier
	q18_7	Le manque de formation adaptée
	q18_8	La paperasserie
	q18_9	La solidarité avec les autres agriculteurs

 ISSN: 2102-6238